

Ten Simple Rules on How to Organise a Bioinformatics Hackathon

Susanne Hollmann^{1,2}, Babette Regierer², Teresa K. Attwood³, Andreas Gisel^{4, #a}, Jacques van Helden⁵,
Gregoire Rossier⁶, Paul J. Kersey^{7, #b}, Eija Korpelainen⁸, Gert Vriend⁹, Erik Bongcam-Rudloff^{10*}

¹Focus Area Plant Genomics and Systems Biology, Institute of Biochemistry and Biology, Potsdam University, Potsdam, DE;

²Leibniz-Institut für Gemüse- und Zierpflanzenbau Großbeeren/Erfurt eV: Grossbeeren, Brandenburg, DE;

³The University of Manchester, Manchester, UK;

⁴Department of Bioinformatics, Institute for Biomedical Technologies, Bari, IT;

⁵Department of Theory and Approaches to Genome, Institut Français de Bioinformatique, Évry, F;

⁶Swiss Institute of Bioinformatics, Lausanne, CH;

⁷EMBL-European Bioinformatics Institute, Hinxton, UK;

⁸IT Center for Science, Espoo, FI;

⁹Centre for Molecular and Biomolecular Informatics, Radboudumc, Nijmegen, NL;

¹⁰Bioinformatics Infrastructure, Swedish University of Agricultural Sciences, Uppsala, SE

^{#a} Current Address: International Institute for Tropical Agriculture, Ibadan, NG;

^{#b} Current Address: Bioinformatics and Genomics, Royal Botanic Gardens, Surrey, UK;

* Corresponding author

E-mail: Erik.Bongcam@slu.se

Abstract

The completion of the human genome sequence triggered world-wide efforts to unravel the secrets sequestered in its deceptively simple code. Numerous bioinformatics projects were undertaken to hunt for genes, to predict their protein products, their functions and post-translational modifications, to analyse protein-protein interactions, etc. In their wake, these diverse projects produced many novel analytic and predictive computer programs, fully optimised for manipulating human genome sequence data. The EU-funded project AllBio “Broadening the Bioinformatics Infrastructure to unicellular, animal and plant science” (FP7 GA 289452), concentrated on non-human genomes, applying human-genome-derived computational solutions to non-human genomics research questions. After several events for community building to get an overview about state-of-the-art of the bioinformatics in the non-human bioinformatics life science areas, the AllBio partners collected “test cases” to identify gaps and existing challenges in the bioinformatics field from the user communities. From the collected test cases, 14 were selected encompassing unicellular organisms, plants and farm animals to be addressed in so-called “hackathons”. A hackathon is an event in which bioinformatics developers convene to work together to produce new software tools and webservice as solutions to the challenges provided. These hackathons have been extremely successful and produced not only new software solutions, but also provided an excellent basis to integrate existing knowledge, to join “wet lab” and “dry lab” experts, to provide hands-on training in bioinformatics, and enforce cross disciplinary interaction. The scientific communities could promote and implement hackathons as a routine activity e.g. in the educational system or for PhD students in the life sciences. Encouraging interdisciplinary interaction in the hackathons could leverage the existing resources and avoid duplication of efforts. The consortium developed 10 Rules as recommendations how to organise hackathons in the life sciences based on the experience in AllBio to make the information available to the life science communities.

Introduction

The completion of the human genome sequence triggered world-wide efforts to unravel the secrets sequestered in its deceptively simple code. Numerous bioinformatics projects were undertaken to

48 hunt for genes, to predict their protein products, their functions and post-translational modifications,
49 to analyse protein-protein interactions, etc. In their wake, these diverse projects produced many novel
50 analytic and predictive computer programs, fully optimised for manipulating human genome
51 sequence data.

52 This focus on exploring the human genome is understandable – we are inevitably anthropocentric. In
53 consequence, considerably less effort has been invested in exploring the many thousands of other
54 available genomes, from unicellular organisms to plants and non-human animals, a detailed
55 understanding of which is, nevertheless, likely to have significant impact on human health and well-
56 being. Nevertheless, recent advances in genome sequencing technology, informatics, automation,
57 artificial intelligence, etc. have allowed us to extend the reach of genomic research to all organisms.
58 Among other large-scale sequencing initiatives for plants, microbes or animals like fish, the Earth
59 Biogenome Project (EBP) [1] was launched with the aim of sequencing the DNA of all life on earth
60 within the next decade. This long-term project will lead both to a greater understanding of Earth's
61 biodiversity and to responsible stewardship of its resources, tackling the most crucial scientific and
62 social challenges of the new millennium. While the focus of the EBP is on the collection of genomic
63 data, other initiatives have centred on data analysis. The EU-funded project AllBio "Broadening the
64 Bioinformatics Infrastructure to unicellular, animal and plant science" (FP7 GA 289452), for example,
65 concentrated on non-human genomes, applying human-genome-derived computational solutions to
66 non-human-genomics research questions [2]. This project involved the collection of a range of
67 biological problems – so-called '**test-cases**' – the computational solutions to some of which were
68 worked out in detail during '**hackathons**'.

69 A hackathon is a short (1-day to 1-week) event where stakeholders with diverse skills and backgrounds
70 gather to develop and implement solutions (usually in the form of software) to relevant problems. The
71 term hackathon is a composite of the words 'hack' (meaning exploratory programming) and
72 'marathon' (a common metaphor for events that are long and intensive). Hackathons are common in
73 informatics communities but still relatively new to the life sciences. In part, this may be because there

are still considerable communication gaps between life- and computational-science researchers. Bioinformatics hackathons, or 'biohackathons' aim to address such gaps specifically by bringing IT professionals (and interested amateurs) and life scientists together to communicate and exchange ideas around practical research questions. They can be very effective, highly productive mechanisms for interdisciplinary teams to work together either to solve well-defined problems or to accelerate solution provision in a particular area. The short time normally available for (bio-) hackathons generally allows for the design and implementation of prototype solutions; if the outputs are to be useful, the developed code must have the potential to undergo subsequent development by interested parties. Therefore, all results should be made available via open accessible platforms so that further improvements can be made after the event has terminated.

The philosophy of the AllBio project was to solicit the needs of life scientists directly, in order to identify interesting challenges. Around 60 such test-cases were collected via questionnaires and interviews, of which 15 (encompassing unicellular organisms, plants, and farm animals) were deemed solvable with adaptations to software or workflows originally designed for working with human-genome data[3]. Eight were subsequently addressed in biohackathons [4,5]. A problem was considered solvable when:

- a generic question relating to the analysis of a unicellular, animal or plant genome had been well defined,
- a community of domain-expert bioscientists and bioinformaticians had been formed, and
- scientific meetings (*in vivo* or *in silico*) had already taken place, and collaborations had begun.

The workflow for AllBio biohackathons involved the collection and selection of the test-cases, preparation and organisation of the events, and finally - in case of success - publication of the results (Fig 1).

Figure 1. Identification of test cases. AllBio workflow illustrating the fate of test-cases proposed by life scientists. After initial interviews, test-cases were collected, and assessed for their tractability. The biohackathon teams comprised the proposer (life scientist), a leader (bioinformatician), 'hackers' (programmers) and, usually, a local organiser. Where a tool or meta-tool

arose from the work, it was proposed for testing during a validation workshop. Ultimately, the team prepared an open source tool, and published or otherwise disseminated the results.

During the AllBio project, a rigorous regime of evaluating past events allowed each biohackathon to build on lessons learned from previous ones. This iterative process demonstrated that, for biohackathons to be successful, the events must be well-prepared, long in advance; the biological problems they set out to tackle must be tractable; they must have access to requisite computational infrastructure, and allow sufficient time to complete the necessary tasks; and they must have efficient leadership, an appropriate mix of skills/expertise, and effective communication strategies. Biohackathons should be preceded by a preparatory phase to check feasibility and practicability (e.g., can the data actually be moved around and read); there should also be commitments afterwards to finalise any tools (or other outputs), to test and validate them with end-users, and to disseminate the results.

Based on the experience gained in AllBio, we present ten rules that we believe are crucial when organising bioinformatics hackathons, or 'BioHackathons': these fall into four main categories - the Problem, the Team, the BioHackathon and the Answer, which are described in detail below. There will, of course, be other important considerations (funding, etc.), but we focus here on the practicalities of organising successful biohackathons.

The Ten Rules

The Problem

Rule 1: Understand the biological problem(s) and select the theme

It might seem self-evident to state that a good starting point is to understand a problem before trying to address it. But solving biological problems via hackathons requires a spectrum of understanding that encompasses the biology of the problem (including *in vivo* aspects), the nature of the data available, the nature of the computational requirements, the nature of the expected output(s), and how all of these can be brought together to implement a viable solution. One of the keys to success is that those responsible for implementing the technical solution(s) must appreciate, at least at some

level, the underlying biology. Ultimately, this requires some investment of time, in order to allow them to begin to understand the language of those whose biological problems they are trying to solve. One way to help achieve this, even for small events, might be to run a small cycle of webinars before the event, in order to give participants more information about the theme. This is likely to facilitate team building, and may also provide opportunities to come up with new ideas for possible approaches and solutions.

Rule 2: Ensure that the problem is tractable

BioHackathons are driven by practical research questions, but not all biological problems are amenable to solution by hackathons. An early step in setting up any such event should therefore be to estimate whether the size of the problem is compatible with the hackathon format. For example, while *de novo* software design is generally not the goal of -hackathons (design of new algorithms tends to require more than just a few days), proof-of-concept implementations can fit the format quite well. Ideally, therefore, the necessary software components must already exist, so that (bio-) hackathon sessions can readily combine them into bespoke workflows. Ideally, workflows should not contain any single point of failure.

Importantly, both the biological data-sets and the software components must be available without restrictions.

The Team

Rule 3: Put together the right team with carefully assigned roles

Start building the team as soon as possible; ideally, aim to establish the core group two months prior to the event. Think about life- and computational-science colleagues and students who have the requisite skills and knowledge in the problem area. Generate a check-list with the minimal requirements needed to ensure that the complete project can be implemented during the event; this will form the basis for participant selection. If necessary, promote the biohackathon widely (*e.g.*, using social media), providing as much information about the event as possible (including when, where, what, how, fees (if needed) and registration forms). To engage biohackathon participants, some

incentives might be helpful: e.g., cooperation with university groups that might be willing to give credit points for participation; or formulating problems whose solutions are suitable for academic publication and crediting those participants as authors.

(Bio-)Hackathon teams are generally most effective when they comprise no more than eight to ten participants. In general, they should include a proposer or ‘biological problem owner’ (typically a life scientist) whose needs will drive the event; a leader (usually a bioinformatician); the ‘hackers’ (bioinformaticians and computer scientists); and, ideally, an overall organiser/coordinator. Those with computational skills should include at least one IT professional or bioinformatician, and programmers with experience in scripting, workflow design, use of ontologies, evaluation of data quality, and so on. These professionals must be able to communicate effectively with the leader and remain focused on the primary objective.

The biohackathon leader is responsible for monitoring and guiding the workflow during the event. The organiser must take responsibility for the overall coordination of the event, maintaining good communication within the team (rule 4), orchestrating the validation (rule 9) and dissemination (rule 10) activities. The organiser must be local to the venue of the (bio-)hackathon, and will be responsible for many mundane practical tasks: reserving the venue, testing bandwidth in the meeting room before the actual hackathon, providing travel instructions, communication with the compute provider, selecting the participants and dealing with subsistence/refreshment issues, etc. Several roles may be assumed by one person, but it is vital that each partner knows his/her role, and that all roles are maintained before, during and after the hackathon itself. To facilitate discussion and assignment of tasks as the project progresses, we suggest adopting a convenient communication platform (*e.g.*, **Trello** [6], **Slack** [7] or comparable platform [8,9,10]).

Rule 4: Communicate effectively and establish the ground rules

Communication – before, during and after hackathons – is key. The value of good communication, and the impact of not getting it right, is hard to over-emphasise. Biohackathons include partners from different disciplines, who tend to speak very different languages – if a biohackathon is to be maximally

productive, it is critical to take time, early on, to identify and resolve potential language barriers. Frequent conversations prior to the biohackathon (in person if possible, or electronically if not), are essential to begin to understand, define and refine the biological question, to identify and shape the overall analytical approach, and thence to build 'ownership' of the tasks. As the technical partners assimilate the nature of the biological problem, and the biological partners begin to appreciate the nature of the technical challenges, the team's purpose, focus and cohesiveness will mature.

If multiple projects are being tackled in one biohackathon, ensure that all requirements have been established beforehand, including the process of team-building, the time-frame available for each problem (equal conditions for every team, so that each has the same relative chance of success), and the rules for allowing participants to move between teams.

Rule 5: Prepare the ground-work well in advance

(Bio-)Hackathons are generally time-limited – good preparatory work is therefore essential. A crucial part of the preparation is to test the necessary software and hardware prior to the event, in order to obviate the occurrence of problems that could reduce the time available for hands-on work. Any heavy computational tasks should be pre-computed to allow participants to hit the ground running with real data. Biohackathon leaders must therefore fully understand all the components in advance, arrange to have them tested in good time, and ensure that both software tools and hardware facilities are adequate for the tasks at hand. For example, CPU-intensive tasks might require massive pre-calculations and/or specialised equipment (such as all-against-all BLAST [11] computations on data-sets with millions of sequences, or the assembly of very large genomes). Just as important is verification of the quality of any data-sets to be used during the event, as poor-quality data-sets are likely to jeopardise the success of (bio-) hackathon sessions. In order not to waste valuable time, any task that can be tackled by a participant in isolation (without requiring the insight of the full team) should be completed in advance. It is vital to test all software and hardware prior to the event. Work with the hackers to establish the hardware requirements. Ensure that hardware equipment/components can be provided or replaced temporarily, if need be.

204 Prepare a budget forecast for the event. The budget will be dedicated to the rental of premises, IT
205 requirements and subsistence. Gather options of suitable venues, and their prices. Look at the
206 premises and find out what the rental includes. Fix the premises for the scheduled date.

207 Decide the total amount you can spend on subsistence. We recommend creating a spreadsheet of all
208 costs. If you have no funds available, you will need to set a fee (which will ultimately be determined
209 by the number of participants, including lecturers, organisers, and so on). If you do have to set fees,
210 you should also be aware of the potential fiscal risks. Involve your administration in the process to
211 ensure that you do not run into trouble: they will know best how to treat fee income. If feasible, search
212 for potential sponsors – *e.g.*, companies with an interest in your BioHackathon theme.

213 We recommend creating a checklist for all tasks to be done before, during and after the event. Spread
214 responsibility between the organisers, but ensure that they do their job seriously. Discuss and agree
215 on the rules and procedures, and take care that they are followed strictly. Fig 2 collates the
216 organisational workflow for a complete biohackathon cycle, including the preparatory,
217 implementation and follow-up phase.

218 **The BioHackathon**

219 **Rule 6: Choose a convenient location**

220 (Bio-)Hackathons should take place at locations that are convenient for the registered number of
221 participants, and that can fulfil all scientific/computing and non-scientific (housing, food, *etc.*) needs.
222 University/national computing centres are likely to offer excellent computational facilities, but may
223 have restrictive opening hours. Hotels, on the other hand, while often very convenient in many
224 aspects, may overestimate the bandwidth they can provide, so this needs to be tested extensively
225 upfront.

226 Specific requirements to consider thus include:

- 227 • location convenient for participants to reach (minimise travel time and cost),
- 228 • short distance between accommodation and meeting venue (if the venue is not the hotel),

- venue technically well equipped (beamer, screen, *etc.*), with liberal opening hours (often, much work is done outside normal working hours, and it is important to facilitate this),
- venue has sufficient and stable bandwidth,
- food and drink are either available at the venue or allowed to be brought in. Often, many productive discussions occur informally over dinner, so arrangements that encourage the participants to keep together while eating are strongly preferred.

Rule 7: Ensure appropriate computer access

All biohackathons are not equal: some will have greater computational requirements than others. Some analyses might run efficiently on participants' own laptops; some might require access to large clusters, supercomputers, dedicated hardware, or the cloud, which universities or national computer centres may be willing to provide. Regardless, the prerequisites are i) fast internet connection at the hackathon venue, and ii) possibility for remote login to the compute facilities before and after the event, so that the groundwork can be prepared beforehand, and any remaining work can be completed later. The local organiser should ensure (and check) that logins are available for all participants, and ideally perform a test-run before the biohackathon. Similarly, if participants use their own laptops, the requisite software should be installed prior to the event. It is recommended to create a Virtual Machine to provide a common compute environment for participants. To gain an overview of the software and hardware that will be needed during the biohackathon, we recommend gathering information about technical requirements via the registration form. Share this information with the hackers at the latest 10 days before the event.

Rule 8: Ensure the duration is sufficient to obtain useful outputs

(Bio-)Hackathons are short, intensive working sessions, typically spanning a few days. Several considerations determine the duration of these events: the complexity of the workflow, how much computer work is envisaged (and how much can be done in advance), the funds available, how much time participants are able to commit, and whether writing documentation and/or article outlines are also intended to be part of the exercise. The expected outputs must therefore be clearly defined early

on, and the duration of the event adjusted accordingly. It generally works well to organise hackathons over a weekend, as this affords participants greater flexibility with their schedules.

To kick off the event, plan to run a series of short lectures to better inform participants about the theme of the BioHackathon, and introduce its biological and computational components. Ensure the availability of suitably qualified lecturers. Disseminate information about these lectures both to the participants and to a broader audience at the latest two weeks before the hackathon. This may stimulate greater interest in the event, and gain visibility within the community.

The lecture hall and work spaces might be at different locations. Ensure that you provide sufficient and detailed information about where and when to go to each place. If there is not sufficient space to comfortably accommodate additional participants in the lecture series, just focus on briefing the team. This can also be done in the form of webinars before the event.

The Answer

Rule 9: Validate the results

BioHackathons aim to address particular biological problems. The events may focus on prototyping ideas, or they may lead to the production of tools or meta-tools that will ultimately be made available to the community. Prior to public release, ‘validation’ events should be organised, in which participants are given opportunities to test the tool(s) with a variety of different data-sets. Even though validation is normally done after hackathons, it should nevertheless be part of the initial planning to ensure that validation data actually exist, and that the software set-up is sufficiently generic to allow its use in validation. In an ideal case, most (if not all) of the original biohackathon team should be present or (remotely) available during validation sessions.

Rule 10: Disseminate the results

Peer-reviewed publications are still the main vehicles for disseminating scientific results, and reusable outputs from biohackathons are a good stimulus for article publication. Public availability of all workflows must also be part of the dissemination strategy. Workflows should therefore be properly

documented and licensed, and inputs and outputs should be appropriately described. Ideally, alongside any publicly accessible documentation or article, small data-sets that can be used by the workflows should also be included. Optionally, Virtual Machine images to run workflows might also be provided. Results should be made available through open accessible platforms (e.g. *seek* [12], *OpenAIRE*, *Zenodo* [13, 14]) that can guarantee longevity, as good workflows that answer biological questions often remain valuable for several years.

Figure 2: Workflow. The scheme demonstrates an optimal workflow for biohackathons, including the preparatory, implementation and follow-up phase for a complete cycle. Each phase is subdivided into different consecutive steps: in particular, the preparatory phase comprises a broad spectrum of tasks, including the selection of challenges, recruiting of participants, organisation of the venue and technical set-up, as well as creation of webinars to prepare participants for the event.

Potential pitfalls

The experience of the AllBio biohackathons provided an inside view of potential pitfalls that might limit the success of such events. A primary challenge is careful selection of appropriate problems - not all are suitable for inclusion in a BioHackathon. It requires expert knowledge from both the biology and bioinformatics fields, to evaluate the challenges and avoid frustration for the participants.

A very specific function that biohackathons can perform is enabling interdisciplinary collaboration between the participants from the different expert fields.. Sufficient time needs to be dedicated to training participants, and finding a common language for discussing the challenges and developing efficient solutions.

Other more practical aspects may limit the success of events: e.g., some early AllBio biohackathons struggled to deliver concrete outputs because:

- their teams were too small (≤ 5 people);
- the team had no real leadership;
- the data-sets on which they were obliged to work were too large to be processed fruitfully within the given time-frame;

- the opening hours of computing centres limited the time available for productive work;
- the distance between hackathon venues and participants' hotels posed time- and cost constraints.

A barrier to success may also occur if the meeting organiser/leader is no longer available after the event; the validation and follow-up phase is essential for summarising the results, and ensuring the quality of solutions that have been developed. Moreover, publication of the results, whether via a journal article or upload to a repository, needs to be completed after the biohackathon. Costs associated with the dissemination of results need to be considered in the overall budget plan.

Conclusion

Biohackathons were powerful tools in the AllBio project for articulating and solving problems in the scientific community [3]. They highlighted the need to take into account the different disciplinary backgrounds of all participants, and hence the vital role of the preparatory phase for ensuring the success of events. They also provided excellent opportunities, especially for young researchers, to learn new skills at the interface between disciplines, to participate in the advancement of their field of research, and to gain unique hands-on training with real challenges.

Some of the rules listed here may seem obvious, trivial, or even superfluous; nevertheless, all proved crucial in real-life scenarios. The ten rules provide practical guidelines for future biohackathon organisers, including preparations before, during and after the event itself.

Acknowledgments

The AllBio project was funded by the EU FP7 work programme KBBE-2011, "FOOD, AGRICULTURE AND FISHERIES, AND BIOTECHNOLOGY", Grant agreement no: 289452. The hackathon team members gratefully acknowledge SARA for providing computer facilities, and SARA for hosting one hackathon. We thank NBIC's BioAssist programme for helping us find excellent programmers and for allowing them to spend adequate time on the projects.

References

- [1] The Earth BioGenome Project (EBP) aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years. Available from: <https://www.earthbiogenome.org/>
- [2] Broadening the Bioinformatics Infrastructure to unicellular, animal, and plant science. Available from: <http://www.allbioinformatics.eu/doku.php/>
<https://cordis.europa.eu/project/rcn/100520/factsheet/en>
- [3] Broadening the bioinformatics infrastructure to unicellular, animal, and plant science D5.3 - Final summary report on solved test cases. Available from: <https://doi.org/10.5281/zenodo.3525052>
- [4] Steinhauf D, Rodriguez A, Vlachakis D, Virgo G, Maksimov V, Kristell C, et al. (2014) Silencing Motifs in the Clr2 Protein from Fission Yeast, *Schizosaccharomyces pombe*. PLoS ONE 9(1): e86948. <https://doi.org/10.1371/journal.pone.0086948>
- [5] Gomez-Cabrero et al.: Data integration in the era of omics: current and future challenges. BMC Systems Biology 20148(Suppl 2):11 doi:10.1186/1752-0509-8-S2-I1
- [6] Online tool of boards, lists, and cards to organise and prioritise projects in a fun, flexible, and rewarding way. Available from: <https://trello.com/>
- [7] Web-based instant messaging service Available from: <https://slack.com/>
- [8] OwnCloud is an open source software for cloud-based collaboration platforms. Available from: <https://owncloud.org/>
- [9] Open source software for cloud-based collaboration platforms. Available from: <https://drive.google.com>
- [10] Dropbox is a modern workspace designed to reduce busywork-so you can focus on the things that matter. Available from: <https://www.dropbox.com>
- [11] Basic Local Alignment Search Tool. Available from: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
- [12] The FAIRDOM Platform uses a stack of software platforms to provide a "Project Commons" – that is, a shared place where projects can organise their own data, models, protocols, workflows in their own managed space. Available from: <https://fair-dom.org/platform/>

- 358 [13] Repository to deposit research products of any type (publication, data, software, other)
- 359 Available from: <https://explore.openaire.eu/participate/deposit-publications>
- 360 [14] Online storage service used for scientific data sets, science-related software, publications,
- 361 reports, presentations, videos, etc. Available from: <https://zenodo.org/>